Statistica Sinica **7**(1997), 1103-1120

SEMIPARAMETRIC METHODS IN LOGISTIC REGRESSION WITH MEASUREMENT ERROR

C. Y. Wang and Suojin Wang

Fred Hutchinson Cancer Research Center and Texas A&M University

Abstract: In this paper we investigate semiparametric estimation methods in logistic regression models with measurement error in the continuous covariates. The measurement error models under consideration have in general two data sets: the validation and nonvalidation data sets. Some covariates are missing in the nonvalidation data set, but a surrogate variable may be available in both data sets. When a covariate variable is missing at random, we consider two kernel assisted estimation methods which extend the pseudo conditional likelihood (PCL) estimate of Breslow and Cain (1988) and the mean-score method of Reilly and Pepe (1995) to continuous covariates and surrogates. The asymptotic results of the two estimators for prospective logistic regression are given. Furthermore, we discuss the asymptotic theory of the PCL estimate in a two-stage case-control (retrospective sampling) study when the covariates and the surrogate are continuous. A simulation study is also given to demonstrate and compare their finite sample properties.

Key words and phrases: Case-control study, errors in variable, kernel smoother, logistic regression, mean-score method, pseudo conditional likelihood.

1. Introduction

Logistic regression is a common tool to investigate factors related to disease incidence. Due to certain reasons such as high cost, some covariate data can only be collected for a small subsample. For example, LDL cholesterol may be related to the risk of heart diseases. However, to reduce the cost, it may be a plausible way to measure the total cholesterol of all study participants and then select a subset to further measure the LDL level. Here the total cholesterol can be treated as a surrogate for LDL. In general, let Y be the binary response, Z be a covariate which is always observed and X be another covariate that may be missing. Assume that W is a surrogate variable for X. We consider the logistic regression model

$$\mathrm{pr}(Y = 1|X, Z) = H(\theta_0 + \theta_1^t X + \theta_2^t Z), \tag{1}$$

where $H(u) = \{1 + \exp(-u)\}^{-1}$ is the logistic distribution function and $\Theta = (\theta_0, \theta_1^t, \theta_2^t)^t$ is a vector of parameters. Let δ_i be a random indicator of X_i being observed. The nonvalidation data set ($\delta_i = 0$) consists of (Y_i, Z_i, W_i) and the

validation data set ($\delta_i = 1$) consists of (Y_i, X_i, Z_i, W_i). In this paper we consider the case where X_i is assumed to be missing at random (MAR, Rubin (1976)) such that the probability of X_i being observed (selection probability) $\text{pr}(\delta_i = 1|Y_i, X_i, Z_i, W_i) = \pi(Y_i, Z_i, W_i)$ depends on (Y_i, Z_i, W_i) but not on X_i. In some cases, the data are obtained in two stages. At the first stage, $(Y_i, Z_i, W_i), i = 1, \ldots, n$ are obtained for all subjects, and at the second stage X_i are measured in the validation data set.

Logistic regression with covariate measurement error models has been an active research area in recent years. Breslow and Cain (1988) proposed a pseudo conditional likelihood (PCL) method for a two-stage case-control study such that at the second stage some X's are observed on each stratum classified by (Y, W) where W is a categorical surrogate. When the missingness of X does not depend on both outcomes and the missing values, Carroll and Wand (1991) and Pepe and Fleming (1991) proposed semiparametric estimators which approximate the likelihood without modeling the distribution of X given (Z, W). Little (1992) reviewed related methods in this field. A mean-score method was proposed by Reilly and Pepe (1995) for discrete covariates when X is MAR. Robins, Rotnitzky and Zhao (1994) proposed efficient estimation by computing an optimal score function in semiparametric models. They showed that the optimal estimator attains the semiparametric variance bound in the sense of Begun, Hall, Huang and Wellner (1983). However, the efficient scores can be computationally challenging, especially when data are continuous.

The methods we investigate in this paper are semiparametric because we do not impose additional models for the nuisance components, such as the selection probabilities of the validation data set or the probability density of $X|(Y, Z, W)$. Therefore, this does not require either submodel assumptions or the use of the EM algorithm (Dempster, Laird and Rubin (1977)). In this paper, we investigate two estimation methods when (Z, W) are continuous, extending the PCL estimator of Breslow and Cain (1988) and the mean-score method of Reilly and Pepe (1995). In Section 2, we extend the PCL estimator for continuous covariates for the prospective sampling scheme such that the first stage sampling is simple random sampling from the source population. Nonparametric kernel estimation is applied to the estimation of selection probabilities and the resulting estimator of Θ is presented. In Section 3, we investigate the mean-score method. The kernel estimator is used to estimate the estimating score when X is not available and the asymptotic distribution result is presented. In Section 4, we discuss two-stage case-control studies and investigate the corresponding asymptotic theory of the PCL estimator (Breslow and Cain (1988)). The result extends the work of Prentice and Pyke (1979) to logistic measurement error models, and it fits the general theory of Carroll, Wang and Wang (1995). A simulation study is conducted in Section 5, and concluding remarks are given in Section 6. All the

technical details for the asymptotic normal theory are provided in Wang and Wang (1996).

To conclude this section, we first note that this paper extends two methods in logistic regression for missing or mismeasured covariate problems to continuous covariates or surrogates. Our estimates of Θ are root-n consistent although the kernel smoothers are not. Note that an application of kernel assisted estimation to partial means in economics was investigated by Newey (1994). Our problem, our approaches and the asymptotic results are different from Newey (1994). In particular, we use different technical linearization tools in the proofs of our main results. In addition, finite sample performance is examined by a simulation study in which we show situations where different estimators are preferred. We also investigate the empirical efficiency of the new methods compared with the efficient parametric maximum likelihood estimator under the additional assumption that the conditional distribution of X given (Y, Z) is perfectly specified.

2. Pseudo Conditional Likelihood Estimate

2.1. Introduction

For notational convenience, let $\mathcal{X} = (1, X^t, Z^t)^t, V = (Z^t, W^t)^t$, $\text{pr}(\delta = 1|Y, V) = \pi(Y, V)$. Recall that the likelihood function (1) is correct when all covariates are observed. As noted in Zhao and Lipsitz (1992), when X is MAR the likelihood in the validation data set is

$$\begin{aligned}\text{pr}(Y = 1|V, X, \delta = 1) &= \frac{\text{pr}(Y = 1|V, X)\text{pr}(\delta = 1|Y = 1, V)}{\sum_{y=0,1} \text{pr}(Y = y|V, X)\text{pr}(\delta = 1|Y = y, V)} \\ &= H\{\Theta^t\mathcal{X} + \log\frac{\pi(1, V)}{\pi(0, V)}\}.\end{aligned}$$

Therefore, conditioning on $\delta = 1$, the maximum likelihood estimator of Θ solves the estimating equation

$$\sum_{i=1}^{n} \delta_i\mathcal{X}_i\Big[Y_i - H\{\Theta^t\mathcal{X}_i + \log\frac{\pi(1, V_i)}{\pi(0, V_i)}\}\Big] = 0. \tag{2}$$

This is similar to the estimation method of Breslow and Cain (1988), but they deal with two-stage case-control data, and with categorical V (see Section 4). Note that (2) contains the selection probability $\pi(Y, V)$. In some cases, $\pi(Y, V)$ in (2) is a nuisance component which is unknown and it remains to be estimated. The approaches of Breslow and Cain (1988) and Schill, Jockel, Drescher and Timm (1993) estimated $\pi(y, v)$ by $\sum_{i=1}^{n} I[Y_i = y, V_i = v, \delta_i = 1]/\sum_{i=1}^{n} I[Y_i = y, V_i = v]$ since their V is categorical.

When V is continuous, natural estimates of $\pi(y, v)$ for $y = 0, 1$ are nonparametric kernel smoothers. Copas (1983) proposed kernel estimates to plot

a binary response against continuous covariates. Here we suggest fitting kernel smoothers of δ's on V's for $Y = 0, 1$ respectively, as estimates of the selection probabilities. Let d be the dimension of V and K be a kernel function of order r. We apply the Nadaraya (1964) and the Watson (1964) estimator such that

$$\widehat{\pi}(y, v) = \frac{\sum_{i \in N(y)} \delta_i K_h(V_i - v)}{\sum_{i \in N(y)} K_h(V_i - v)}, \tag{3}$$

where $N(y) = \{i : Y_i = y\}, K_h(\cdot) = K(\cdot/h)$ and h is the bandwidth. Optimal bandwidth conditions for h are given next.

2.2. Preliminary properties

Define $H_i = H(\Theta^t \mathcal{X}_i), H_{*i} = H[\Theta^t \mathcal{X}_i + \log\{\pi(1, V_i)/\pi(0, V_i)\}]$, $\widehat{H}_{*i} = H[\Theta^t \mathcal{X}_i + \log\{\widehat{\pi}(1, V_i)/\widehat{\pi}(0, V_i)\}]$. The estimating score of the pseudo conditional likelihood estimator $\widehat{\Theta}_{PCL}$ is

$$U_n(\Theta, \widehat{\pi}) = n^{-1/2} \sum_{i=1}^{n} \delta_i \mathcal{X}_i (Y_i - \widehat{H}_{*i}).$$

By direct calculations, we have the following lemma which is used frequently in the proofs of the main results.

Lemma 1. *Let* $\overline{H}_i = 1 - H_i$ *and* $\overline{H}_{*i} = 1 - H_{*i}$. *Then* $\pi(0, V_i) H_{*i} \overline{H}_i = \pi(1, V_i) \overline{H}_{*i} H_i$.

Define $S_i(y) = \mathcal{X}_i(y - H_{*i})$. When π is known, to see that $U_n(\Theta, \pi)$ is an unbiased estimating score, we note that if we denote the distribution of (X, V) by $F(X, V)$ then

$$\begin{aligned} E\{\delta_i S_i(Y_i)\} &= E[E\{\delta_i \mathcal{X}_i (Y_i - H_{*i}) | Y_i, V_i, X_i\}] = E\{\pi(Y_i, V_i) \mathcal{X}_i (Y_i - H_{*i})\} \\ &= \int (1, z^t, x^t)^t [\pi(1, v) \overline{H}_* H + \pi(0, v)\{-H_*\} \overline{H}] dF_{X,V}(x, v) = 0. \end{aligned}$$

The last equation follows from Lemma 1. Note that the above expectations hold at the true parameter Θ. For notational convenience, let $\eta_n = \{nh^{2r} + 1/(nh^{2d})\}^{1/2}$. We assume the following conditions.

(A1) For $y = 0, 1$ $g(y, v)$ has the rth continuous derivative *a.e.*, and for $k = 1, \ldots, r$, $|\dfrac{\partial^k}{\partial v^k} g(y, v)|$ is bounded *a.e.*

(A2) For any given compact set $\mathcal{C}$ in the domain of v, there exists $c_1 > 0$ such that the selection probabilities $\pi(y, v) \geq c_1$ for $y = 0, 1$ and all $v \in \mathcal{C}$.

(A3) For $y = 0, 1$ $\pi(y, v)$ has rth continuous derivative *a.e.*, and for $k = 1, \ldots, r$, $|\dfrac{\partial^k}{\partial v^k} \pi(y, v)|$ is bounded *a.e.*

(A4) $E(\mathcal{X}\mathcal{X}^t)$ exists and $E\{\mathcal{X}\mathcal{X}^t\pi(0,V)\overline{H}H_*\}$ is positive definite.

First, we establish the consistency of $\widehat{\Theta}_{PCL}$.

Lemma 2. *Under Conditions* (A1)-(A4), *if* $nh^{2r} \to 0$ *and* $nh^{2d} \to \infty$, *then there exists a solution* $\widehat{\Theta}_{PCL}$ *to* $U_n(\Theta, \widehat{\pi}) = 0$ *with probability converging to* 1 *as* $n \to \infty$, *and* $\widehat{\Theta}_{PCL} \to \Theta$ *in probability.*

The asymptotic theory is based mainly on the linear approximation

$$U_n(\Theta, \widehat{\pi}) = n^{-1/2} \sum_{i=1}^{n} \Big[\delta_i S_i(Y_i) + (-1)^{Y_i}\{\delta_i - \pi(Y_i, V_i)\}\mathcal{L}_i(Y_i)\Big] + O_p(\eta_n),$$

where $\mathcal{L}_1(Y_1)$ is defined in (5) below. Since the main terms have zero expectations, under conditions (A1)-(A4), the result follows. Next, we briefly illustrate how we need the rate of order η_n. We note that a kernel smoother has the rate of bias of order h^r and that of $U_n(\Theta, \widehat{\pi}) - U_n(\Theta, \pi)$ is $n^{1/2}h^r$. The error term in linearizing $U_n(\Theta, \widehat{\pi})$ is of order $O_p\{n^{1/2}(\widehat{\pi} - \pi)^2\}$, which is $O_p(n^{1/2}h^{2r} + n^{-1/2}h^{-d})$.

2.3. The limit distribution

Let $g(y, v)$ denote the probability density/mass function of (Y, V). Under the moment condition (A4), we define $H_{*i}^{(1)} = H_{*i}\overline{H}_{*i}$, $\Omega_i(y) = E\{\mathcal{X}_i H_{*i}^{(1)} | Y_i = y, V_i\}$,

$$G(\Theta, \pi) = E\{\mathcal{X}_1\mathcal{X}_1^t\pi(0, V_1)\overline{H}_1 H_{*1}\}, \tag{4}$$

$$\mathcal{L}_i(y) = \Omega_i(y) + \frac{\pi(1-y, V_i)g(1-y, V_i)}{\pi(y, V_i)g(y, V_i)}\Omega_i(1-y), \tag{5}$$

$$M_1(\Theta, \pi) = E\Big[(-1)^{Y_1} S_1(Y_1)\pi(Y_1, V_1)\{1 - \pi(Y_1, V_1)\}\mathcal{L}_1^t(Y_1)\Big],$$

$$M_2(\Theta, \pi) = E\Big[\pi(Y_1, V_1)\{1 - \pi(Y_1, V_1)\}\mathcal{L}_1(Y_1)\mathcal{L}_1^t(Y_1)\Big],$$

$$M(\Theta, \pi) = M_1(\Theta, \pi) + M_1^t(\Theta, \pi) + M_2(\Theta, \pi). \tag{6}$$

We now present the limit distribution for the PCL estimator.

Theorem 1. *Let* $\widehat{\Theta}_{PCL}$ *be the semiparametric estimate of* Θ *solving* $U_n(\Theta, \widehat{\pi}) = 0$, *where* $\widehat{\pi}$ *is a kernel smoother of* π *given in* (3) *assuming that the bandwidth* h *satisfies* $nh^{2d} \to \infty$ *and* $nh^{2r} \to 0$ *and the boundary kernel is applied at the boundary points. Then under Conditions* (A1)-(A4), $n^{1/2}(\widehat{\Theta}_{PCL} - \Theta)$ *has an asymptotic normal distribution with mean zero and asymptotic covariance matrix*

$$G^{-1}(\Theta, \pi)\{G(\Theta, \pi) + M(\Theta, \pi)\}G^{-t}(\Theta, \pi). \tag{7}$$

A detailed proof of this theorem can be found in Wang and Wang (1996). We now describe the covariance estimates. Define $G_n(\widehat{\Theta}_{PCL}, \widehat{\pi}) =$

$n^{-1}\sum_{i=1}^n \delta_i \mathcal{X}_i \mathcal{X}_i^t \widehat{H}_{*i}^{(1)}(\widehat{\Theta}_{PCL})$,

$$\widehat{H}_{*i}(\widehat{\Theta}_{PCL}) = H\Big\{\widehat{\Theta}_{PCL}^t \mathcal{X}_i + \log\frac{\widehat{\pi}(1,V_i)}{\widehat{\pi}(0,V_i)}\Big\}, \quad \widehat{S}_i(y) = \mathcal{X}_i\{y - \widehat{H}_{*i}(\widehat{\Theta}_{PCL})\},$$

$$\widehat{\Omega}_i(y) = \frac{\sum_{k\in N(y)} \dfrac{\delta_k}{\widehat{\pi}(Y_i,V_i)} \mathcal{X}_k \widehat{H}_{*k}^{(1)}(\widehat{\Theta}_{PCL}) K_h(V_k - V_i)}{\sum_{k\in N(y)} K_h(V_k - V_i)},$$

$$\widehat{\mathcal{L}}_i(y) = \widehat{\Omega}_i(y) + \frac{\widehat{\pi}(1-y,V_i)\widehat{g}(1-y,V_i)}{\widehat{\pi}(y,V_i)\widehat{g}(y,V_i)}\widehat{\Omega}_i(1-y),$$

$$\widehat{g}(y,V_i) = \frac{1}{nh^d}\sum_{j\in N(y)} K_h(V_j - V_i),$$

$$\widehat{M}_1(\widehat{\Theta}_{PCL},\widehat{\pi}) = n^{-1}\sum_{i=1}^n \Big[(-1)^{Y_i}\widehat{S}_i(Y_i)\widehat{\pi}(Y_i,V_i)\{1-\widehat{\pi}(Y_i,V_i)\}\widehat{\mathcal{L}}_i^t(Y_i)\Big],$$

$$\widehat{M}_2(\widehat{\Theta}_{PCL},\widehat{\pi}) = n^{-1}\sum_{i=1}^n \Big[\widehat{\pi}(Y_i,V_i)\{1-\widehat{\pi}(Y_i,V_i)\}\widehat{\mathcal{L}}_i(Y_i)\widehat{\mathcal{L}}_i^t(Y_i)\Big],$$

$$\widehat{M}(\widehat{\Theta}_{PCL},\widehat{\pi}) = \widehat{M}_1(\widehat{\Theta}_{PCL},\widehat{\pi}) + \widehat{M}_1^t(\widehat{\Theta}_{PCL},\widehat{\pi}) + \widehat{M}_2(\widehat{\Theta}_{PCL},\widehat{\pi}).$$

Then a consistent estimator of the covariance matrix of $n^{1/2}(\widehat{\Theta}_{PCL} - \Theta)$ is

$$G_n^{-1}(\widehat{\Theta}_{PCL},\widehat{\pi})\{G_n(\widehat{\Theta}_{PCL},\widehat{\pi}) + \widehat{M}(\widehat{\Theta}_{PCL},\widehat{\pi})\}G_n^{-t}(\widehat{\Theta}_{PCL},\widehat{\pi}).$$

3. Mean-Score Method

3.1. Introduction

When the missingness process does not depend on outcome, Carroll and Wand (1991) and Pepe and Fleming (1991) proposed semiparametric estimates of Θ which estimate the likelihood function without modeling the conditional distribution of X given V. Their estimates may be inconsistent when the selection of X depends on the response Y. When V is discrete, Reilly and Pepe (1995) generalized Pepe and Fleming's (1991) method and proposed a mean-score method for the case that the probability of $\delta_i = 1$ depends on both Y_i and V_i, i.e., X is MAR. To see their method, first assume that the conditional density of X given (Y,V) in the validation is known, then an unbiased estimating equation is

$$\Phi_n(\Theta) = n^{-1/2}\sum_{i=1}^n \Big[\delta_i\phi(Y_i,X_i,Z_i) + (1-\delta_i)E\{\phi(Y_i,X_i,Z_i)|Y_i,V_i,\delta_i=1\}\Big] = 0,$$

where ϕ is the estimating score when all data are observed. In this paper, we concentrate on the case of logistic regression and hence $\phi(Y_i,X_i,V_i) = \mathcal{X}_i\{Y_i -$

$H(\Theta^t\mathcal{X}_i)\}$. The method can be easily extended to other situations. The unbiasedness of $\Phi_n(\Theta)$ follows because $E\{\delta_i\phi(Y_i,X_i,Z_i)\} = E[\pi(Y_i,V_i)E\{\phi(Y_i,X_i,Z_i)|Y_i,V_i,\delta_i=1\}]$ and $E[E\{\phi(Y_i,X_i,Z_i)|Y_i,V_i,\delta_i=1\}]=0$. Note that $E\{\phi(Y_i,X_i,Z_i)|Y_i,V_i,\delta_i=1\} = E\{\phi(Y_i,X_i,Z_i)|Y_i,V_i\}$ since X_i is independent of δ_i given (Y_i,V_i) for X_i is MAR. Because the conditional density of X given (Y,V) is unknown in general, when V is discrete, Reilly and Pepe suggested using $\widehat{E}\{\phi(Y_i,X_i,Z_i)|Y_i,V_i,\delta_i=1\}$ which takes averages of $\phi(Y_k,X_k,Z_k)$ on the validation data such that $Y_k=Y_i$ and $V_k=V_i$.

When V is continuous, we suggest estimating the conditionally expected estimating score $E\{\phi(Y_i,X_i,Z_i)|Y_i,V_i\}$ in the nonvalidation data set by using a kernel smoother based on the validation data. Therefore, the corresponding estimating equation of the mean-score method is

$$\begin{aligned}&\widehat{\Phi}_n(\Theta)\\ &= n^{-1/2}\sum_{i=1}^{n}\Big[\delta_i\phi(Y_i,X_i,Z_i)+(1-\delta_i)\frac{\sum_{k\in N(Y_i)}\delta_k\phi(Y_k,X_k,Z_k)K_h(V_k-V_i)}{\sum_{k\in N(Y_i)}\delta_k K_h(V_k-V_i)}\Big]\\ &= 0\end{aligned} \tag{8}$$

with the solution $\widehat{\Theta}_{MS}$ as the mean-score estimate. The basic idea of the method is to use the score $\phi(Y_i,X_i,Z_i)$ when the data are complete, and use the estimated score when X_i is missing or mismeasured. Similar idea and calculations may be extended to surrogate endpoint problems (Pepe, Reilly and Fleming (1994)). The limit distribution of $\widehat{\Theta}_{MS}$ is given next.

3.2. The limit distribution

We assume the following conditions.

(B1) For $y=0,1$ and $\delta=0,1$, $f_{V|Y=y,\delta}(v)$ has the rth continuous derivative *a.e.*, and for $k=1,\ldots,r$, $|\frac{\partial^k}{\partial v^k}f_{V|Y=y,\delta}(v)|$ is bounded *a.e.*

(B2) $E(\mathcal{X}_1\mathcal{X}_1^tH_1^{(1)})$ exists and is positive definite.

(B3) Let $\mathcal{A}(Y,V) = E\{\mathcal{X}_1\mathcal{X}_1^tH_1^{(1)}|(Y,V)\}$. For $y=0,1$, $\mathcal{A}(y,v)$ has the rth continuous derivative *a.e.*, and for k=1,…, r, $|\frac{\partial^k}{\partial v^k}\mathcal{A}(y,v)|$ is bounded *a.e.*

The following lemma shows the consistency of $\widehat{\Theta}_{MS}$.

Lemma 3. *Under Conditions* (B1)-(B3), *if* $nh^{2r}\to 0$ *and* $nh^{2d}\to\infty$, *then there exists a solution* $\widehat{\Theta}_{MS}$ *to* $\widehat{\Phi}_n(\Theta)=0$ *with probability converging to* 1 *as* $n\to\infty$, *and* $\widehat{\Theta}_{MS}\to\Theta$ *in probability.*

Similar to Lemma 2, the asymptotic distribution is based on the linearization

$$\widehat{\Phi}_n(\Theta) = n^{-1/2}\sum_{i=1}^{n}\Big[\delta_i\phi_i+(1-\delta_i)\{E(\phi_i|Y_i,V_i)+T_{ni}\}\Big]+O_p(\eta_n),$$

where $T_{ni} = \{n_{VY_i} h^d f_{V|Y=Y_i,\delta=1}(V_i)\}^{-1} \sum_{k\in N(Y_i)} \delta_k R_{ki} K_h(V_k - V_i)$. For $j = 0, 1$ define $n_{Vj} = \sum_{k\in N(j)} \delta_k$ as the total number of observations in the validation data set such that $Y_k = j$, $n_{Pj} = \sum_{k\in N(j)}(1 - \delta_k)$ as the total number of observations in the nonvalidation data set such that $Y_k = j$. Assume that for $j = 0, 1$, $n_{Pj}/n_{Vj} \to \tau_j$ with $\tau_j < \infty$. Let $R_{ki} = \phi_k - E(\phi_i|Y_i, V_i)$ and $R^*_{kk} = R_{kk}\{f_{V|Y=Y_k,\delta=0}(V_k)\}\{f_{V|Y=Y_k,\delta=1}(V_k)\}^{-1}$. Under the moment condition (B2), we define $\Psi(\Theta) = E[\mathcal{X}_1 \mathcal{X}_1^t H_1^{(1)}]$, $J_1(\Theta) = E\{\pi(Y_1, V_1)(\phi_1 + \tau_{Y_1} R^*_{11})(\phi_1 + \tau_{Y_1} R^*_{11})^t\}$, $J_2(\Theta) = E[\{1 - \pi(Y_1, V_1)\}E(\phi_1|Y_1, V_1)\ E^t(\phi_1|Y_1, V_1)]$, and $J(\Theta) = J_1(\Theta) + J_2(\Theta)$. We now present the asymptotic distribution of the mean-score estimate for continuous covariates.

Theorem 2. *Let the mean-score estimate $\widehat{\Theta}_{MS}$ be the semiparametric estimate of Θ solving $\widehat{\Phi}_n(\Theta) = 0$ with the bandwidth h satisfying $nh^{2d} \to \infty$ and $nh^{2r} \to 0$ and the boundary kernel is applied at the boundary points. Then under Conditions* (B1)-(B3), $n^{1/2}(\widehat{\Theta}_{MS} - \Theta)$ *has an asymptotic normal distribution with mean zero and asymptotic covariance matrix $\Psi^{-1}(\Theta)J(\Theta)\Psi^{-t}(\Theta)$.*

A detailed proof of this theorem can be found in Wang and Wang (1996).

3.3. Covariance estimation

By some direct calculations, one can show that $\Psi_n(\Theta) \to \Psi(\Theta)$ *in prob.*, where

$$\Psi_n(\Theta) = n^{-1} \sum_{i=1}^{n} \Big[\delta_i \mathcal{X}_i \mathcal{X}_i^t H_i^{(1)} + (1 - \delta_i) \frac{\sum_{k\in N(Y_i)} \delta_k \mathcal{X}_i \mathcal{X}_i^t H_i^{(1)} K_h(V_k - V_i)}{\sum_{k\in N(Y_i)} \delta_k K_h(V_k - V_i)}\Big].$$

Hence, by Lemma 3, we have $\Psi_n(\widehat{\Theta}_{MS}) \to \Psi(\Theta)$ *in prob.* Furthermore,

$$\begin{aligned}\widehat{J}_1(\widehat{\Theta}_{MS}) &= n^{-1} \sum_{i=1}^{n} \delta_i \{\phi_i(\widehat{\Theta}_{MS}) + \frac{n_{PY_i}}{n_{VY_i}} \widehat{R}^*_{ii}\}\{\phi_i(\widehat{\Theta}_{MS}) + \frac{n_{PY_i}}{n_{VY_i}} \widehat{R}^*_{ii}\}^t \\ &\to J_1(\Theta) \text{ in prob.},\end{aligned}$$

where $\widehat{R}^*_{ii} = [\phi_i(\widehat{\Theta}_{MS}) - \widehat{E}\{\phi_i(\widehat{\Theta}_{MS})|Y_i, V_i\}]\ \widehat{f}_{V|Y=Y_i,\delta=0}\ (V_i)\widehat{f}^{-1}_{V|Y=Y_i,\delta=1}(V_i)$, $\phi_i(\widehat{\Theta}_{MS}) = \mathcal{X}_i\{Y_i - H(\widehat{\Theta}^t_{MS}\mathcal{X}_i)\}$, $\widehat{E}\{\phi_i(\widehat{\Theta}_{MS})|Y_i, V_i\} = \{\sum_{k\in N(Y_i)} \delta_k \phi_k(\widehat{\Theta}_{MS})$ $K_h(V_k - V_i)\}\{\sum_{k\in N(Y_i)} \delta_k K_h(V_k - V_i)\}^{-1}$, and $\widehat{f}_{V|Y=Y_i,\delta=0}$ and $\widehat{f}_{V|Y=Y_i,\delta=1}$ are the kernel density estimators of V given the response Y_i in the validation and nonvalidation data, respectively. Finally,

$$\begin{aligned}\widehat{J}_2(\widehat{\Theta}_{MS}) &= n^{-1} \sum_{i=1}^{n} (1 - \delta_i)\widehat{E}\{\phi_i(\widehat{\Theta}_{MS})|Y_i, V_i\}[\widehat{E}\{\phi_i(\widehat{\Theta}_{MS})|Y_i, V_i\}]^t \\ &\to J_2(\Theta) \text{ in prob.},\end{aligned}$$

leading to a consistent estimator $\Psi_n^{-1}(\widehat{\Theta}_{MS})\widehat{J}(\widehat{\Theta}_{MS})\Psi_n^{-t}(\widehat{\Theta}_{MS})$ for the asymptotic covariance matrix $\mathrm{cov}\{n^{1/2}(\widehat{\Theta}_{MS} - \Theta)\}$, where $\widehat{J}(\widehat{\Theta}_{MS}) = \widehat{J}_1(\widehat{\Theta}_{MS}) + \widehat{J}_2(\widehat{\Theta}_{MS})$.

4. Two-Stage Case-Control Studies

In this section, we investigate the PCL estimator in a two-stage case-control study.

4.1. Introduction

A case-control (retrospective) study is different from a cohort (prospective) study in the sampling design. Breslow and Cain (1988) investigated a two-stage case-control study when V is categorical. They considered the case where $V = W$, and the results in their paper may be extended to the setting such that $V = (Z^t, W^t)^t$. At the first stage all V's are observed for n_0 controls and n_1 cases, and at the second stage partial X's are selected and observed from each stratum ($V = 1, \ldots, L_V$ for some L_V). Their approach is similar to the maximum pseudo conditional likelihood of Hsieh, Manski and McFadden (1985). Schill, Jockel, Drescher and Timm (1993) generalized the method of Prentice and Pyke (1979) to the sampling scheme considered by Breslow and Cain (1988). They maximized two likelihood components jointly, whereas Breslow and Cain maximized a likelihood which contains estimated nuisance parameters. We consider continuous V and therefore the second stage data are based on the selection probabilities $\pi(y, v)$. The major difference between this section and previous sections in this paper is that the first stage sampling is retrospective and hence fixed n_0 controls and n_1 cases are obtained in each subpopulation.

4.2. Estimation equations

When there is no measurement error, Prentice and Pyke (1979) showed that the retrospective likelihood of (Z, X) given Y is

$$\mathrm{pr}^*(z, x|Y = y) = H^y(\theta_0^* + \theta_1^t x + \theta_2^t z)(1 - H(\theta_0^* + \theta_1^t x + \theta_2^t z))^{1-y} dP^*_{X,Z}(x, z)\frac{n}{n_y},$$

where $P^*_{X,Z}(x, z)$ is the joint distribution function of (X, Z) under case–control sampling and θ_0^* is a new intercept. Note that θ_0 is not identifiable (Prentice and Pyke (1979)).

The estimating equation (2) in the two-stage case-control sampling is the same except for the intercept. Therefore we define $\Theta = (\theta_0^*, \theta_1^t, \theta_2^t)^t$ in this section and $S_i(y) = \mathcal{X}_i(y - H_{*i})$, where $H_{*i} = H[\Theta^t\mathcal{X}_i + \log\{\pi(1, V_i)/\pi(0, V_i)\}]$ as before. Recall that we defined $U_n(\Theta, \pi) = n^{-1/2}\sum_{i=1}^n \delta_i S_i(Y_i)$. For the retrospective

sampling, the unbiasedness holds since

$$
\begin{aligned}
&E\Big\{\sum_{i=1}^{n}\delta_i S_i(Y_i)|\tilde{Y}\Big\} = E\Big[E[\sum_{i=1}^{n}\delta_i\mathcal{X}_i\{Y_i - H_{*i}\}|\tilde{Y},\tilde{V}]|\tilde{Y}\Big]\\
&= E\Big[E[\sum_{i=1}^{n}\pi(Y_i,V_i)\mathcal{X}_i\{Y_i - H_{*i}\}|\tilde{Y},\tilde{V}]|\tilde{Y}\Big]\\
&= \int\sum_{y=0}^{1} n_y\pi(y,v)(1,x^t,z^t)^t(y-H_*)\frac{n}{n_y}H^y(\theta_0^*+\theta_1^t x+\theta_2^t z)\{\overline{H}(\theta_0^*+\theta_1^t x+\theta_2^t z)\}^{1-y}\\
&\qquad\qquad dP^*_{X,V}(x,v)\\
&= n\int(1,x^t,z^t)^t[\pi(0,v)\{-H_*\}\overline{H}(\theta_0^*+\theta_1 x+\theta_2 z)+\pi(1,v)\{\overline{H}_*\}H(\theta_0^*+\theta_1 x+\theta_2 z)]\\
&\qquad\qquad dP^*_{X,V}(x,v) = 0,
\end{aligned}
$$

where $\tilde{Y} = (Y_1,\ldots,Y_n)^t$, $\tilde{V} = (V_1,\ldots,V_n)^t$, $P^*_{X,V}(x,v)$ is the marginal distribution of (X,V) under the retrospective sampling scheme, H_* in the above calculations is defined as H_{*i} with intercept θ_0^* and $(X_i,V_i) = (x,v)$. The last equation holds by Lemma 1.

4.3. Covariance estimation

The derivation of the asymptotic covariances is slightly different from that of the usual prospective sampling. Denote $[\cdots]$ as a repeat of the terms in brackets immediately proceeding. First we note that

$$
\begin{aligned}
&n^{-1}\mathrm{cov}\{\sum_{i=1}^{n}\delta_i S_i(Y_i)|\tilde{Y}\}\\
&= \frac{n_0}{n}\mathrm{cov}[\delta_1\mathcal{X}_1\{-H_{*1}\}|Y_1=0] + \frac{n_1}{n}\mathrm{cov}[\delta_1\mathcal{X}_1\{1-H_{*1}\}|Y_1=1]\\
&= \frac{n_0}{n}E[\pi(0,V_1)\mathcal{X}_1\mathcal{X}_1^t H_{*1}^2|Y_1=0] + \frac{n_1}{n}E[\pi(1,V_1)\mathcal{X}_1\mathcal{X}_1^t\{1-H_{*1}\}^2|Y_1=1]\\
&\quad -\frac{n_0}{n}[E\{\pi(0,V_1)\mathcal{X}_1 H_{*1}|Y_1{=}0\}][\cdots]^t - \frac{n_1}{n}[E\{\pi(1,V_1)\mathcal{X}_1(1-H_{*1})|Y_1{=}1\}][\cdots]^t\\
&= G(\Theta,\pi) - (\frac{n}{n_0}+\frac{n}{n_1})C_1C_1^t + o(1),
\end{aligned}
$$

where $G(\Theta,\pi)$ is defined in (4) with intercept θ_0^* and $C_1 = \int(1,x^t,z^t)^t\pi(0,v)$ $H_*\overline{H}dP^*_{X,V}(x,v)$. The last equation follows from Lemma 1.

With some algebra, it can be shown that the retrospective covariance of $n^{1/2}(\widehat{\Theta}_{PCL}-\Theta)$ is

$$
G^{-1}(\Theta,\pi)\{G(\Theta,\pi)+M(\Theta,\pi)-(\frac{n}{n_0}+\frac{n}{n_1})C_1C_1^t\}G^{-t}(\Theta,\pi), \tag{9}
$$

where $M(\Theta,\pi)$ is defined in (6).

Comparing the retrospective formula (9) and the prospective formula (7) and using the fact that C_1 is the first column of $G(\Theta, \pi)$, we therefore have shown that the prospective asymptotic covariance is asymptotically correct for $\widehat{\theta}_1$ and $\widehat{\theta}_2$ for the two-stage case-control sampling design. This extends the results of Prentice and Pyke (1979). As a final remark of this section, we note that the result fits the general theory of Carroll, Wang and Wang (1995) because the second stage samples are not stratified on V. This is different from the case with discrete V, where inference is conditioned on $(\tilde{Y}, \tilde{V})$. Nevertheless, from the result for continuous V, it appears that the prospective covariance formula is slightly *conservative* for the biased sampling design of Breslow and Cain (1988) when V is discrete.

5. A Simulation Study

In this section, we demonstrate numerical performance of the two estimators developed in this paper. The first one is a semiparametric PCL estimator which extends the method of Breslow and Cain since π is estimated by smoothers. The second is the smoothing mean-score method, which extends the estimate of Reilly and Pepe (1995) to continuous covariates. We also list the complete-case analysis and the conditional likelihood estimator of Θ when true π is applied. The complete-case analysis ($\widehat{\Theta}_{CC}$) applies the usual logistic regression to the validation data $\{(Y_i, X_i) : \delta_i = 1, \text{ for } i \in \{1, \ldots, n\}\}$ only. The conditional likelihood estimator ($\widehat{\Theta}_{CL}$) is the same as the PCL estimator except for using true π.

In this simulation study, we consider the cases where $n = 200$ and $n = 500$, respectively. The covariates X's were generated from a uniform $[-1, 1]$ distribution and $W = X + \sigma U$, where U is from a uniform $[-1, 1]$ distribution which is independent of X, and $\sigma = .2$, $\sigma = 1$ were used in Tables 1 and 2, respectively. The binary response Y was generated by the model $\text{pr}(Y = 1|X) = H(\theta_0 + \theta_1 X)$, where $\theta_0 = .5$ and $\theta_1 = 1$. In this study, W is the surrogate for X. The validation data indicator δ_i in Tables 1 and 2 was generated by ths selection probability such that $\text{pr}(\delta_i = 1|Y_i, W_i) = \{1 + \exp(-Y_i - W_i)\}^{-1}$. On average, about 63% of the observations are validation data in which X was observed. The PCL estimator was obtained by applying the uniform kernel function and the bandwidth $h = 2\widehat{\sigma}_{y,W} n^{-1/3}$ and $h = 4\widehat{\sigma}_{y,W} n^{-1/3}$, respectively, to estimate $\pi(y, v)$ for $y = 0$ and $y = 1$, where $\widehat{\sigma}_{y,W}$ is the sample standard deviation of W for $y = 0$ and $y = 1$. The mean-score estimate was obtained by solving (8). The bandwidth selection satisfies the conditions in Theorems 1 and 2 and is a convenient choice. In analyzing real data, it may be helpful to apply some bandwidth criteria such as the generalized cross validation or the approximate asymptotic mean integrated square error. We used the local linear smoother

(Fan (1993)) since it seems to have better finite sample performance. Also, we set $\widehat{\pi}$ bounded above a small positive number, say .01. The estimated means and standard errors were obtained from 500 independent runs. In our experience this selection works reasonably well numerically. Note that the complete-case analysis which uses only the validation data has large bias since the selection probabilities depend on (Y, W). In the tables "s.e. of $\widehat{\theta}$" denotes the actual s.e. approximation using the repeated runs while "mean(s.e.)" is the average value of repeated estimated s.e. from the derived formulas.

Table 1. Simulation study: X is MAR and the measurement error is small.

				$h = 2\widehat{\sigma}_{y,W} n^{-1/3}$		$h = 4\widehat{\sigma}_{y,W} n^{-1/3}$	
		$\widehat{\Theta}_{CC}$	$\widehat{\Theta}_{CL}$	$\widehat{\Theta}_{PCL}$	$\widehat{\Theta}_{MS}$	$\widehat{\Theta}_{PCL}$	$\widehat{\Theta}_{MS}$
n=200	$\widehat{\theta}_0 - \theta_0$	.417	.030	.022	.020	.022	.016
	(s.e. of $\widehat{\theta}_0$'s)	(.206)	(.206)	(.163)	(.157)	(.161)	(.156)
	(mean(s.e.))	(.208)	(.208)	(.164)	(.153)	(.161)	(.151)
	95% cov. prob.	.472	.946	.952	.942	.956	.944
	$\widehat{\theta}_1 - \theta_1$	−.204	.017	−.011	−.011	.003	−.057
	(s.e. of $\widehat{\theta}_1$'s)	(.387)	(.388)	(.294)	(.286)	(.303)	(.262)
	(mean(s.e.))	(.374)	(.378)	(.318)	(.267)	(.314)	(.248)
	95% cov. prob.	.902	.954	.976	.948	.966	.934
n=500	$\widehat{\theta}_0 - \theta_0$	.399	.012	.005	.002	.007	.000
	(s.e. of $\widehat{\theta}_0$'s)	(.132)	(.132)	(.098)	(.096)	(.098)	(.096)
	(mean(s.e.))	(.130)	(.130)	(.101)	(.096)	(.100)	(.095)
	95% cov. prob.	.136	.958	.966	.956	.954	.956
	$\widehat{\theta}_1 - \theta_1$	−.212	.009	−.003	−.005	−.002	−.004
	(s.e. of $\widehat{\theta}_1$'s)	(.227)	(.228)	(.179)	(.167)	(.181)	(.160)
	(mean(s.e.))	(.233)	(.234)	(.193)	(.166)	(.192)	(.159)
	95% cov. prob.	.842	.964	.964	.942	.956	.940

Note. Simulation study when $\Theta = (.5, 1)^t$. The true $\pi = \{1 + \exp(-Y - W)\}^{-1}$, and on average there are 63% subjects in the validation set. The surrogates follow the model $W = X + \sigma U$, where X and U are independent and uniformly distributed in $[-1, 1]$, and $\sigma = .2$. The notation $\widehat{\Theta}_{CC}$ is for the estimate of the complete-case analysis, $\widehat{\Theta}_{PCL}$ for the pseudo conditional likelihood estimate (using estimated $\widehat{\pi}$), $\widehat{\Theta}_{CL}$ for the conditional likelihood estimate (using true π), and $\widehat{\Theta}_{MS}$ for the mean-score estimate. There were 500 replications.

Table 2. Simulation study: X is MAR and the measurement error is large.

				$h = 2\widehat{\sigma}_{y,W}n^{-1/3}$		$h = 4\widehat{\sigma}_{y,W}n^{-1/3}$	
		$\widehat{\Theta}_{CC}$	$\widehat{\Theta}_{CL}$	$\widehat{\Theta}_{PCL}$	$\widehat{\Theta}_{MS}$	$\widehat{\Theta}_{PCL}$	$\widehat{\Theta}_{MS}$
n=200	$\widehat{\theta}_0 - \theta_0$	.394	.027	.010	.043	.025	.030
	(s.e. of $\widehat{\theta}_0$'s)	(.209)	(.208)	(.167)	(.180)	(.164)	(.178)
	(mean(s.e.))	(.208)	(.208)	(.169)	(.175)	(.166)	(.171)
	95% cov. prob.	.506	.946	.956	.940	.958	.946
	$\widehat{\theta}_1 - \theta_1$	−.174	.018	−.019	.000	.003	−.006
	(s.e. of $\widehat{\theta}_1$'s)	(.384)	(.385)	(.352)	(.392)	(.348)	(.375)
	(mean(s.e.))	(.375)	(.378)	(.361)	(.395)	(.357)	(.376)
	95% cov. prob.	.910	.952	.958	.966	.964	.956
n=500	$\widehat{\theta}_0 - \theta_0$	.379	.013	.004	.012	.010	.005
	(s.e. of $\widehat{\theta}_0$'s)	(.133)	(.133)	(.109)	(.114)	(.107)	(.104)
	(mean(s.e.))	(.130)	(.130)	(.104)	(.106)	(.103)	(.102)
	95% cov. prob.	.162	.940	.948	.952	.946	.956
	$\widehat{\theta}_1 - \theta_1$	−.191	.001	−.013	.003	−.007	−.013
	(s.e. of $\widehat{\theta}_1$'s)	(.240)	(.241)	(.220)	(.248)	(.220)	(.219)
	(mean(s.e.))	(.233)	(.234)	(.220)	(.272)	(.219)	(.247)
	95% cov. prob.	.852	.944	.950	.976	.944	.972

Note. Simulation study when $\Theta = (.5, 1)^t$. The true $\pi = \{1 + \exp(-Y - W)\}^{-1}$, and on average there are 63% subjects in the validation set. The surrogates follow the model $W = X + \sigma U$, where X and U are independent and uniformly distributed in $[-1, 1]$, and $\sigma = 1$. The notation $\widehat{\Theta}_{CC}$ is for the estimate of the complete-case analysis, $\widehat{\Theta}_{PCL}$ for the pseudo conditional likelihood estimate (using estimated $\widehat{\pi}$), $\widehat{\Theta}_{CL}$ for the conditional likelihood estimate (using true π), and $\widehat{\Theta}_{MS}$ for the mean-score estimate. There were 500 replications.

The data in Table 3 were generated similarly to that of Table 1 except that the validation set was selected by $\text{pr}(\delta_i = 1|Y_i, W_i) = \{1 + \exp(-W_i)\}^{-1}$. There were about 50% subjects with X available. The complete-case analysis in this case is still valid because the missingness does not depend on the outcome. However, it is not as efficient as the other two semiparametric methods. Note that $\widehat{\Theta}_{CC} = \widehat{\Theta}_{CL}$ in this case because the "offset" $\log(\pi(1, V)/\pi(0, V)) = 0$.

Table 3. Simulation study: X is MAR and not depending on Y.

				$h = 2\widehat{\sigma}_{y,W} n^{-1/3}$		$h = 4\widehat{\sigma}_{y,W} n^{-1/3}$	
		$\widehat{\Theta}_{CC}$	$\widehat{\Theta}_{CL}$	$\widehat{\Theta}_{PCL}$	$\widehat{\Theta}_{MS}$	$\widehat{\Theta}_{PCL}$	$\widehat{\Theta}_{MS}$
n=200	$\widehat{\theta}_0 - \theta_0$	.003	.003	−.011	−.012	.015	.015
	(s.e. of $\widehat{\theta}_0$'s)	(.242)	(.242)	(.162)	(.170)	(.172)	(.163)
	(mean(s.e.))	(.224)	(.224)	(.166)	(.159)	(.163)	(.153)
	95% cov. prob.	.940	.944	.960	.946	.930	.934
	$\widehat{\theta}_1 - \theta_1$	.013	.013	−.019	.009	.018	−.043
	(s.e. of $\widehat{\theta}_1$'s)	(.412)	(.412)	(.309)	(.349)	(.332)	(.271)
	(mean(s.e.))	(.408)	(.408)	(.319)	(.295)	(.313)	(.259)
	95% cov. prob.	.946	.946	.962	.940	.950	.934
n=500	$\widehat{\theta}_0 - \theta_0$	.010	.010	.005	.003	.001	−.002
	(s.e. of $\widehat{\theta}_0$'s)	(.144)	(.144)	(.096)	(.094)	(.103)	(.098)
	(mean(s.e.))	(.140)	(.140)	(.103)	(.097)	(.101)	(.096)
	95% cov. prob.	.952	.952	.966	.958	.944	.942
	$\widehat{\theta}_1 - \theta_1$	.011	.011	.000	−.003	−.012	−.040
	(s.e. of $\widehat{\theta}_1$'s)	(.264)	(.264)	(.194)	(.187)	(.187)	(.162)
	(mean(s.e.))	(.253)	(.253)	(.194)	(.179)	(.190)	(.167)
	95% cov. prob.	.952	.952	.956	.940	.962	.946

Note. Simulation study when $\Theta = (.5, 1)^t$. The true $\pi = \{1 + \exp(-W)\}^{-1}$, and on average there are 50% subjects in the validation set. The surrogates follow the model $W = X + \sigma U$, where X and U are independent and uniformly distributed in $[-1, 1]$, and $\sigma = .2$. The notation $\widehat{\Theta}_{CC}$ is for the estimate of the complete-case analysis, $\widehat{\Theta}_{PCL}$ for the pseudo conditional likelihood estimate (using estimated $\widehat{\pi}$), $\widehat{\Theta}_{CL}$ for the conditional likelihood estimate (using true π), and $\widehat{\Theta}_{MS}$ for the mean-score estimate. There were 500 replications.

One might be interested to see the relative efficiency (RE) of the proposed estimators compared to the information bound under the semiparametric models (Bickel, Klaassen, Ritov and Wellner (1993), Chapter 4; Robins, Rotnitzky and Zhao (1994)). One difficulty is that the computations seem to be challenging given that the covariate data are continuous. Alternatively, we have compared the maximum likelihood (ML) estimator under an additional model assumption. As in Blackhurst and Schluchter (1989), we firstly generated Z from uniform $[-1, 1]$, then generated Y such that $\text{pr}(Y = 1|Z) = .5$. We finally generated X from normal distribution with mean $\lambda^2 Y + Z - (1 + .5\lambda^2)$ and variance λ^2. Note that under this data generating device Y given (X, Z) follows (1) with $\Theta = (1, 1, -1)$. We considered the missing data problem here and no surrogate (W)

was available. The validation data set was determined by the selection probability that $\text{pr}(\delta_i = 1|Y_i, Z_i) = \{1 + \exp(-Y_i - Z_i)\}^{-1}$. It has about 60% of all the subjects. The ML estimator based on the above model can be obtained without intensive numerical calculations such as the EM algorithm. By convention we define the RE of the PCL estimator with respect to the ML estimator by the ratio of the variance of the ML estimator over that of the PCL estimator. The bandwidth $h = 4\widehat{\sigma}_{y,W} n^{-1/3}$ was used. Figure 1 shows the curves of the RE of the PCL estimator with respect to the ML estimator over the range of λ from .2 to 2 for $n = 500$ based on 500 replications. The curves are for θ_0, θ_1 and θ_2 respectively. We did the RE for 10 equal spaced λ's first and then estimated the RE curves by using kernel smoothers. When λ is small, the RE is close to 1 since X values are almost determined by given (Y, Z) values. The RE generally decreases as λ increases as we would expect. Nevertheless, the RE of the PCL estimator remains higher than 80% except for $\lambda = 2$. When $\lambda = 2$, it stands for the situation where X given Z is the mixture of two quite separated normals and that is probably the situation in which the semiparametric methods need to pay a high price for not using the information in the strong additional submodel assumption of X given Z or X given (Y, Z). In addition, we have also studied the empirical RE of $\widehat{\Theta}_{MS}$, which is very similar to that of $\widehat{\Theta}_{PCL}$. Note that since the ML estimator is efficient in a subclass of the semiparametric models (as in Robins, Rotnitzky and Zhao (1994)), the RE would be expected to be higher when the comparison is made with the semiparametric information bound.

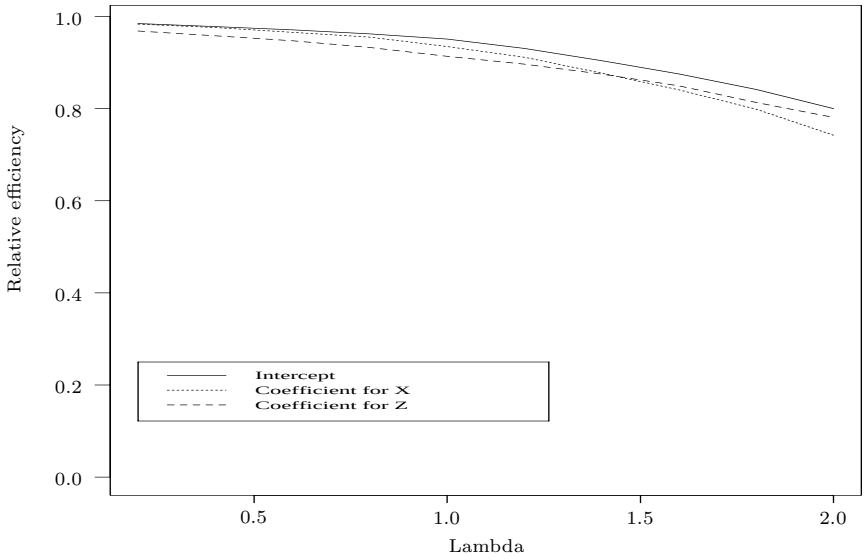

Figure 1. The efficiency of $\widehat{\Theta}_{PCL}$ relative to the parametric MLE

From the empirical results, we have the following summary:

(i) Both bandwidth selections $h = 2\hat{\sigma}_{y,W}n^{-1/3}$ and $h = 4\hat{\sigma}_{y,W}n^{-1/3}$ perform equally well for $\widehat{\Theta}_{PCL}$. However, for $\widehat{\Theta}_{MS}$, using $h = 4\hat{\sigma}_{y,W}n^{-1/3}$ seems to yield slightly higher efficiency.

(ii) $\widehat{\Theta}_{PCL}$ is more efficient than $\widehat{\Theta}_{CL}$. This phenomenon is quite usual, namely, plugging in estimated selection probabilities is better than using the true values even if we know the true values.

(iii) When measurement error is small ($\sigma = .2$), $\widehat{\Theta}_{MS}$ has higher efficiency; while $\widehat{\Theta}_{PCL}$ is superior when the measurement error is large ($\sigma = 1$) and $n = 200$. They have about the same efficiency when $n = 500$ and $\sigma = 1$.

(iv) When X given (Y, Z) is modeled perfectly, the semiparametric estimator $\widehat{\Theta}_{PCL}$ still remains competitive with reasonable RE with respect to the ML estimator.

6. Concluding Remarks

We have proposed two methods for both prospective and retrospective logistic regression when covariates are continuous. They are not restricted to two-stage design. For example, they are useful for missing covariates problems also. The PCL method is useful only for logistic regression, while the MS estimator can be applied to general semiparametric models. The bandwidth conditions require that the order r of the kernel function K is greater than the dimension d of (Z, W). For the pseudo conditional likelihood estimator, one solution to this problem is to assume that the selection probability π depends only on a linear combination $\mathcal{B} = \gamma_1 Y + \gamma_2^t V$ for some (γ_1, γ_2^t). For the mean-score method, we may consider the extension of the dimension reduction methodology of Carroll, Knickerbocker and Wang (1995). More generally, we may consider the generalized additive model (Hastie and Tibshirani (1986)), or the generalized partially linear single-index models (Carroll, Fan, Gijbels and Wand (1995)) to estimate the selection probabilities (π_i) and the mean scores ($E\{\phi|Y_i, V_i\}$). Nevertheless, parametric modeling of the unknown conditional distributions may also be a sensible approach when there is a large number of potentially relevant covariates.

In addition to the methods we described, simple imputation in logistic regression has been proposed in the literature. For example, Rosner, Willett and Spiegelman (1989) and Sepanski, Knickerbocker and Carroll (1994) considered the case when the validation data set is random. When X is MAR, in addition to the methods described in the previous sections, we may apply weighted estimating equations (see, for example, Flanders and Greenland (1991), Zhao and Lipsitz (1992)). Wang, Wang, Zhao and Ou (1997) extend it to continuous covariates. To gain efficiency, the smoothing techniques of this paper may be extended to Robins, Rotnitzky and Zhao (1994) in principle, but the calculations

and computations may be very intensive. On the other hand, the findings from our simulations show that the relative efficiency of the PCL or MS estimator is reasonably high compared to the ML estimator given the data are perfectly modeled. In summary, the proposed semiparametric methods appear to be versatile, likely to be of use in practice.

Acknowledgement

The research of C. Y. Wang was supported by the National Cancer Institute grant CA 53996. The research of S. Wang was supported by National Science Foundation grant DMS-9504589, National Security Agency grant MDA904-96-1-0029, and Texas A&M University's Scholarly and Creative Activities Program grant 95-59. We are grateful to Raymond Carroll, the Editor and a referee for valuable comments.

References

Begun, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11**, 432-452.

Bickel, P., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore: Johns Hopkins University Press.

Blackhurst, D. W. and Schluchter, M. D. (1989). Logistic regression with a partially observed covariate. *Comm. Statist. Ser. A* **18**, 163-177.

Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two–stage case–control data. *Biometrika* **75**, 11-20.

Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *J. Roy. Statist. Soc. Ser. B* **53**, 573-585.

Carroll, R. J., Fan, J., Gijbels, I and Wand, M. (1995). Generalized partially linear single-index models. Preprint.

Carroll, R. J., Knickerbocker, R. K. and Wang, C. Y. (1995). Dimension reduction in a semiparametric regression model with errors in covariates. *Ann. Statist.* **23**, 161-181.

Carroll, R. J., Wang, S. and Wang, C. Y. (1995). Prospective analysis of logistic case-control studies. *J. Amer. Statist. Assoc.* **90**, 157-169.

Copas, J. B. (1983). Plotting p against x. *Appl. Statist.* **32**, 25-31.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196-216.

Flanders, W. D. and Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statist. Med.* **10**, 739-747.

Hastie, T. J. and Tibshirani, R. J. (1986). Generalized additive models. *Statist. Sci.* **1**, 297-318.

Hsieh, D. A., Manski, C. F. and McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations. *J. Amer. Statist. Assoc.* **80**, 651-662.

Little, R. J. A. (1992). Regression with missing X's: A Review. *J. Amer. Statist. Assoc.* **87**, 1227-1237.

Nadaraya, E. A. (1964). On non-parametric estimates of density functions and regression curves. *Theory Probab. Appl.* **10**, 186-190.

Newey, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory* **10**, 233-253.

Pepe, M. S. and Fleming, T. R. (1991). A nonparametric method for dealing with mismeasured covariate data. *J. Amer. Statist. Assoc.* **86**, 108-113.

Pepe, M. S., Reilly, M. and Fleming, T. R. (1994). Auxiliary outcome data and the mean score method. *J. Statist. Plann. Inference* **42**, 137-160.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case–control studies. *Biometrika* **66**, 403-411.

Reilly, M. and Pepe, M. S. (1995). A meanscore method for missing and auxiliary covariate data in regression models. *Biometrika* **82**, 299-314.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846-866.

Rosner, B., Willett, W. C. and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statist. Med.* **8**, 1051-1069.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

Schill, W., Jöckel, K.-H., Drescher, K. and Timm, J. (1993). Logistic analysis in case–control studies under validation sampling. *Biometrika* **80**, 339-352.

Sepanski, J. H., Knickerbocker, R. K. and Carroll, R. J. (1994). A semiparametric correction for attenuation. *J. Amer. Statist. Assoc.* **89**, 1366-1373.

Wang, C. Y. and Wang, S. (1996). Asymptotic theory of semiparametric methods in logistic regression with measurement error. Technical report #251, Department of Statistics, Texas A & M University.

Wang, C. Y., Wang, S., Zhao, L. P. and Ou, S. T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *J. Amer. Statist. Assoc.* **92**, 512-525.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya A* **26**, 359-372.

Zhao, L. P. and Lipsitz, S. (1992). Designs and analysis of two–stage studies. *Statist. Med.* **11**, 769-782.

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, MP 1002, Seattle, WA 98104, USA.

E-mail: cywang@fhcrc.org

Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA.

E-mail: sjwang@stat.tamu.edu

(Received September 1995; accepted September 1996)